On the Efficiency of Image Metrics for Evaluating the Visual Quality of 3D Models

Guillaume Lavoué, *Senior, IEEE,* Mohamed Chaker Larabi, *Senior, IEEE,* and Libor Váša

Abstract—3D meshes are deployed in a wide range of application processes (e.g. transmission, compression, simplification, watermarking and so on) which inevitably introduce geometric distortions that may alter the visual quality of the rendered data. Hence, efficient model-based perceptual metrics, operating on the geometry of the meshes being compared, have been recently introduced to control and predict these visual artifacts. However, since the 3D models are ultimately visualized on 2D screens, it seems legitimate to use images of the models (i.e. snapshots from different viewpoints) to evaluate their visual fidelity. In this work we investigate the use of image metrics to assess the visual quality of 3D models. For this goal, we conduct a wide-ranging study involving several 2D metrics, rendering algorithms, lighting conditions and pooling algorithms, as well as several mean opinion score databases. The collected data allow (1) to determine the best set of parameters to use for this image-based quality assessment approach and (2) to compare this approach to the best performing model-based metrics and determine for which use-case they are respectively adapted. We conclude by exploring several applications that illustrate the benefits of image-based quality assessment.

Index Terms—Computer Graphics, Image Quality Assessment, 3D Mesh Visual Quality Assessment, Perceptual metrics.

1 Introduction

THREE-dimensional graphical data, commonly represented using triangular meshes, are now commonplace in many fields of industry including digital entertainment, mechanical engineering, cultural heritage, scientific visualization, medical imaging and architecture. Moreover, use of 3D data is bound to increase for the general public with the proliferation of intuitive 3D sculpting and modeling tools, affordable 3D printers and community model repositories. 3D data will probably also play a significant role in the future evolution of the Web with the development of Web3D technologies (WebGL, X3D and so on). As a result of this increasing and heterogeneous use, 3D meshes are deployed in a wide range of application processes which include transmission, compression, simplification, remeshing, filtering, watermarking and so on. These operations inevitably introduce artefacts which often alter the visual quality of the rendered data.

In order to deliver satisfactory Quality of Experience to application users, it is critical to be able to evaluate the quality of the distorted 3D data, i.e. the degree of annoyance caused by the artifacts. Simple geometric metrics, such as Hausdorff distance and root mean squared error (RMS), are only weakly correlated with human vision, hence mesh visual quality (MVQ) metrics have been recently studied by the scientific community [1], [2]; their goal is to predict the perceived visual fidelity of distorted 3D data with respect to the original. These metrics operate on the geometry of the meshes being compared; they are referred to as *model-based* metrics in the rest of the paper.

Since the 3D models are ultimately visualized on 2D screens, 2D snapshots of the models can also be considered to evaluate their visual fidelity, in the manner of Lindstrom and Turk [3] for driving simplification. Such *image-based* approaches could be particularly efficient since many successful image quality assessment (IQA) metrics have been introduced in the last decade [4]. This paper attempts to answer the following questions: Which is the best method for predicting 3D mesh visual fidelity? image-based or model-based metrics?

Ten years ago, Rogowitz and Rushmeier [5] and Watson et al. [6] attempted to answer this question, followed more recently by Cleju and Saupe [7]. While these prior works provide interesting insights regarding this question, they are limited to the evaluation of a single type of distortion (only simplification), and test a very small number of metrics (mostly the simplest ones: Hausdorff, 2D/3D mean and root mean squared distances) and, more generally, they only consider a small number of variables in their protocols (i.e. lighting condition, shaders, number of views, etc). These limitations lead to somewhat contradictory conclusions.

Our goal is to conduct a complete and comprehensive study that aims to provide conclusive results regarding this issue. This study is complex to carry out since a large number of parameters are involved: which 2D metrics offer optimum performance? how to choose the 2D views of the 3D models to use in these metrics? how to combine the quality scores from different views into a single one for the 3D model? Hence, our study involves a large number of variables: we have considered 7 image

- *G. Lavoué is with the University of Lyon, CNRS, LIRIS UMR 5205, France*
  E-mail: glavoue@liris.cnrs.fr.
- *M.C. Larabi is with the University of Poitiers, CNRS, XLIM-SIC UMR 7252, France.*
- *Libor Váša is with the University of West Bohemia, Pilsen, Czech Republic.*

metrics, 2 rendering algorithms, 4 lighting conditions, 5 ways of combining image metric results and 4 existing 3D object databases with mean opinion scores for the evaluation. Note that since this study concerns the evaluation of geometric artefacts only, we consider simple rendering styles without texture or complex shader. In total we have generated and analyzed around 60,000 images for this study. We have conducted a statistical analysis of the collected results allowing us to determine the best set of parameters and image metrics to use for image-based quality assessment. We then compare this approach to the best performing model-based metrics and determine for which use-case they are respectively best suited. Finally, we explore several applications where image-based quality assessment is particularly well-suited.

The rest of this paper is organized as follows: section 2 describes related work about visual quality assessment of 3D graphics. Then, sections 3 and 4 present, respectively, the image metrics and the rendering protocols that we consider in our study. Section 5 provides information on the subjective mean opinion score databases used for our evaluation and section 6 details the results of our study. Finally, section 7 illustrates benefits and applications of image-based metrics.

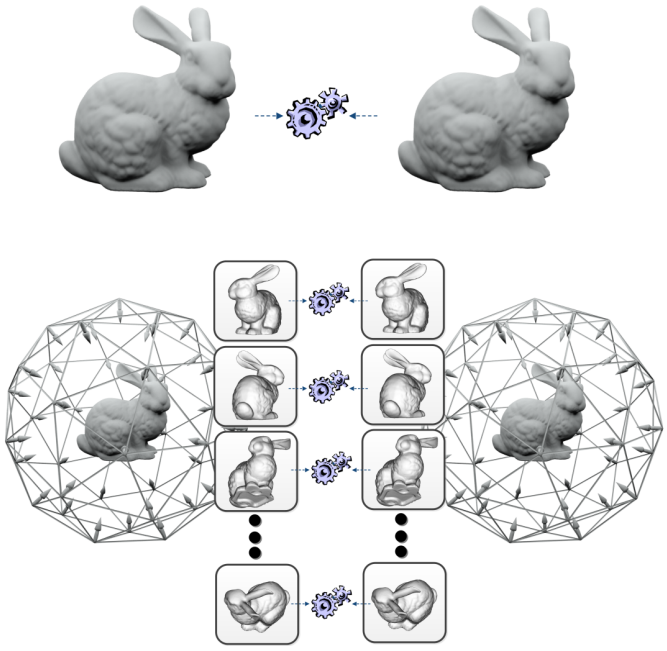

Fig. 1. Illustration of model-based quality metrics operating directly on geometry (top) and image-based quality metrics operating on 2D snapshots (bottom).

2 Related Work on Visual Quality Assessment of Graphical Data

In the field of 2D image processing, the research into objective image quality assessment metrics is highly developed [4]. Pioneer works (VDP [8], Sarnoff VDM [9]) were introduced twenty years ago and hundreds of metrics have been proposed since then, such as the often-cited SSIM (Structural SIMilarity) index, introduced by Wang and Bovik [10] which exploits an important aspect of human visual system (HVS) perception linked to structural information. With a more theoretical definition, the VIF [11] has been developed with the aim to quantify loss of image information due to the distortion process. More recently, the HDR-VDP2 metric [12] offers a visual difference predictor suited for any range of luminance. We refer the reader to section 3 which describes some of the most relevant image metrics, used in our study, in detail.

In comparison with this field of image quality assessment, the visual quality of 3D shapes has as yet been far less investigated. Two kinds of approaches are available for this task: model-based and image-based approaches. Model-based approaches operate directly on the geometry and/or texture of the meshes being compared, while image-based approaches considers rendered images of the 3D models.

2.1 Model-based quality metrics

In the field of computer graphics, the first attempts to evaluate the visual fidelity of 3D objects were simple geometric distances, mainly used for driving mesh simplification [13]. Cignoni et al. [14] provided the metro tool with an implementation of *maximum* (i.e. Hausdorff), *root mean squared* (RMS) and *mean* geometric distances between 3D models. However, these simple geometric measurements represent a very poor predictor of visual fidelity, as demonstrated in several studies [1], [15]. Hence, researchers have introduced perceptually-motivated metrics. These full reference metrics compare the distorted and original 3D models and compute a score reflecting visual fidelity. For example, Karni and Gotsman [16] proposed combining the RMS geometric distance between corresponding vertices with the RMS distance of their Laplacian coordinates (which reflect the degree of surface smoothness)). Lavoué [17] and Torkhani et al. [18] proposed metrics based on local differences in curvature statistics, while Váša and Rus [15] considered the dihedral angle differences. These metrics consider local variations in attribute values at vertex or edge level, which are then pooled into a global score. In contrast, Corsini et al. [19] and Wang et al. [20] compute one global roughness value per 3D model and then derive a simple global roughness difference. For the roughness calculation, they consider, respectively, dihedral angles and variance of the geometric Laplacian [19] and the Laplacian of the Gaussian curvature [20]. Some authors also proposed quality assessment metrics for textured 3D meshes [21], [22], which integrate both texture and geometry information. A very recent review [2] details these works and compares their performance regarding their correlation with mean opinion scores derived from subjective rating experiments; this study

shows that MSDM2 [17], FMPD [20] and DAME [15] are excellent predictors of visual quality.

2.2 Image-based quality metrics

Apart from these model-based quality metrics, operating on 3D geometry, many researchers have used 2D image metrics to evaluate the visual quality of 3D models, mainly in the context of simplification and LoD management for rendering. Bolin and Meyer [23] introduced a simplified version of the Sarnoff VDM [9] to optimize sampling for ray-tracing algorithms. Qu and Meyer [24] considered the visual masking properties of 2D texture maps to drive simplification and remeshing of textured meshes. They evaluate the potential masking effect of surface signals (textures, bump maps, etc) using the 2D Sarnoff VDM [9]. Some authors considered 2D models of the contrast sensitivity function (CSF) [25], [26], [27] to drive simplification or LoD selection. Another example is Zhu et al. [28], who studied the relationship between the viewing distance and the perceptibility of model details using image metrics (VDP [8] and SSIM [10]) for optimal design of discrete LoD for visualization of complex 3D building facades. While most of these works are view-dependent, 2D image metrics can be used in a view independent way by evaluating them on a set of snapshots of the 3D objects, taken from different viewpoints. This kind of image-based view-independent approach was considered by Lindstrom and Turk [3], who evaluate the impact of simplification using a fast image quality metric (RMS error) computed on snapshots taken from 20 different camera positions regularly sampled on a bounding sphere.

2.3 Comparing the two approaches

An important question for the future of 3D graphics quality assessment and, by extension, for the whole computer graphics community is: *what is the best way to evaluate the visual quality of 3D graphical objects, image-based or model-based metrics?* As pointed out in [3], the main benefit of using image-based metrics to evaluate the visual quality of 3D objects is that the complex interactions between the various properties involved in appearance (geometry, texture, normals) are naturally handled, avoiding the problem of how to combine and weigh them. On the other hand, Rogowitz and Rushmeier [5] advocate model-based metrics since they show that 2D judgments do not provide a good predictor of 3D object quality, implying that the quality of 3D objects cannot be correctly predicted by the quality of static 2D projections. To demonstrate this, the authors have conducted two subjective rating experiments; in the first, the observers rated the quality of 2D static images of simplified 3D objects, while in the second they rated an animated sequence of these images, showing a rotation of the 3D objects. Results show that (1) the lighting conditions strongly influence the perceived quality and (2) the observers perceive differently the quality of the 3D objects according to whether they observe still images or animations. Watson et al. [6] also compared the performance of several image-based (Bolin-Meyer [23] and Mean Squared Error) and model-based (mean, max and RMS) metrics. They conducted several subjective experiments to study the visual fidelity of simplified 3D objects, including perceived quality rating. Their results showed a good performance of 2D metrics (Bolin-Meyer [23] and MSE) as well as the mean 3D geometric distance as predictor of the perceived quality. The main limitation of this study is that the authors only considered a single view of the 3D models. More recently, Cleju and Saupe [7] designed another subjective experiment for evaluating the perceived visual quality of simplified 3D models and found that image-based metrics generally perform better than model-based metrics. In particular, they found that 2D mean squared error and SSIM provide good results, whereas the performance of SSIM is more sensitive to the 3D model type. For model-based metrics, like Watson et al. [6], they showed that the mean geometric distance performs better than RMS, which is better than Hausdorff.

While the works mentioned above provide interesting insights into the relative accuracy of image-based and model-based metrics, they possess some serious limitations:

- Only a single type of distortion is evaluated/rated (only simplification).
- As the image-based and model-based metrics compared are quite simple, a lot of new and more efficient works have been introduced since.
- Accuracy of visual quality prediction is evaluated based on the results of a single subjective experiment.
- A limited number of variables are tested, whereas many parameters may influence the image-based results (number of views, rendering algorithm, method of combining results for each view, lighting conditions).

Hence, we present a comprehensive study that compares the results of the most recent and efficient image metrics (see section 3) with the best model-based metrics (presented above). In order to take into account all the parameters involved we consider different lighting conditions (section 4.1), rendering protocols (section 4.2), and different ways of combining the image metric values (section 4.3). Metrics performance is evaluated on several databases (section 5) containing different 3D models subject to different types of distortions, and associated with mean opinion scores.

3 Image Quality Metrics

Over the last decade, the field of image quality assessment has been extremely active and hundreds of

related works can be found. This is due to the scientific community's interest in acquiring tools for accurate measurement of the fidelity of a given algorithm.
With the aim of conducting the present study, we did our utmost to select metrics either known for their efficiency or for their widespread use in the community. To date, several reviews, surveys and chapters have been published regarding image quality metrics for different application fields [29], [30], [31], [32], [33]. When browsing the literature, it can be noticed that some metrics, namely MSE, PSNR, SSIM [10], VIF [11], are often used as anchors to demonstrate the efficiency of a proposed metric. For this study, we opted for signal-based, structural similarity, feature similarity, information fidelity metrics in addition to a perceptually weighted metric and a visual difference predictor.

3.1 Signal-based image metrics: MSE and PSNR

The Peak Signal-to-Noise Ratio (PSNR) is undoubtedly the most widely used metric to date even with the advent of an impressive number of metrics. It owes its popularity mainly to the simplicity of its description, its ease of understanding and low complexity. PSNR, which originates from the signal processing community, aims at assessing fidelity thanks to the ratio between the maximum possible power of a signal and the power of corrupting noise affecting it.

PSNR is expressed in terms of the logarithmic decibel (dB) scale based on the mean square error (MSE) as described below.

$$PSNR = 10\ log_{10} \frac{I_{max}^2}{MSE} \tag{1}$$

where I_{max} is the maximum possible value of a pixel (255 for 8-bit images).

There is no standard formulation of PSNR for multi-channel (*e.g.* RGB) or multi-view (*e.g.* stereoscopic) images. Several authors use either an average of MSEs over all channels / views or an average of individual PSNRs.

Typical values of PSNR for lossy compression are between 30 and $50\ dB$ where higher values indicate better fidelity. However, care must be taken when interpreting results because their validity greatly depends on the content, the compared algorithms and the spatial or spectral distribution of noise [34]. Therefore, perceptual validity of PSNR is very disputable and several studies demonstrated its inefficiency on specific impairments.

3.2 MS-SSIM: Multi-Scale Structural Similarity Index Metric

The Structural Similarity Index Metric (SSIM) proposed by Wang et al. [10] assumed an important position in the quality evaluation community and beyond, thanks to the very interesting tradeoff between complexity and correlation with human judgment. However, it is quite difficult to achieve user unanimity when dealing with this metric. Indeed, it is somehow difficult to interpret results when two impaired images are close in terms of quality leading to similar conclusions regarding PSNR as to usability. Moreover, as reported by the authors themselves, SSIM is a single-scale metric while the viewing conditions are variable. To cope with this limitation, an extension of this metric called Multi-Scale SSIM (MS-SSIM) has been proposed [35]. The extension inherits all the features introduced in the single-scale version. To avoid redundancy, we focused here only on the extended version of the structural similarity.

The MS-SSIM metric takes as input the reference and impaired images and compares two features called contrast c and structure s defined by:

$$c(I, I') = \frac{2\sigma_I \sigma_{I'} + c_1}{\sigma_I^2 + \sigma_{I'}^2 + c_1}, \tag{2}$$

$$s(I, I') = \frac{\sigma_{II'} + c_2}{\sigma_I + \sigma_{I'} + c_2}, \tag{3}$$

where σ_* and σ_{**} represent the variance and the covariance of luminance, respectively. c_* are constants used for computation stability.

This scale 1 processing is iterated at every scale and moves from scale to scale are performed by applying a low-pass filter and downsampling the filtered image by a factor of 2 until scale M. While contrast and structure are computed at each scale, another feature called luminance l is computed only on the smallest scale (M) as described below:

$$l_M(I, I') = \frac{2\mu_I \mu_{I'} + c_3}{\mu_I^2 + \mu_{I'}^2 + c_3}. \tag{4}$$

where μ_* is the luminance mean. The MS-SSIM score is obtained by computing and combining the aforementioned features on local image patches i at different scales j, as described by equation 5.

$$MS{-}SSIM_i = [l_M(I_i, I_i')]^{\alpha_M} \prod_{j=1}^{M} [c(I_{j,i}, I'_{j,i})]^{\beta_j} [s(I_{j,i}, I'_{j,i})]^{\gamma_j} \tag{5}$$

Exponents α_M, β_j and γ_j are used to adjust the relative importance of the different features. The weighting values are obtained by means of a psycho-physical study conducted with a panel of ten observers. These local scores are then averaged into a single score per image.

3.3 FSIM: Feature Similarity Index Metric

Zhang et al. proposed to exploit two important features of the HVS, namely phase congruency (PC) and gradient magnitude (GM) [36]. The former, PC, originates from the idea that features are considered as noticeable at points where the phase is maximal for Fourier components. This is confirmed by physiological and psychophysical studies on how the mammalian visual system detects and identifies salient features in an image. The latter feature, *i.e. GM*, is used to cope with the fact that PC is contrast invariant while the HVS is sensitive

to image local contrast and this important feature has to be taken into account.

Based on the above description, Zhang et al. formulated their proposed metric, based on phase congruency and gradient magnitude, as given by equation 6.

$$FSIM = \frac{\sum_i S_i \cdot PC_i^{max}}{\sum_i PC_i^{max}} \tag{6}$$

where $\sum_i$ represents a summation over local image patches i; $PC_i^{max} = max(PC_{I_i}, PC_{I'_i})$ and $S_i = [S_{PC_i}]^{\alpha}[S_{GM_i}]^{\beta}$ is the weighted combination of PC and GM similarities between the original image patch I_i and the impaired one I'_i. While α and β may be used to give more importance to one feature or another, the author chose to give them equal importance. Similarity between PCs (S_{PC}) and GMs (S_{GM}) is expressed as follows:

$$S_{PC} = \frac{2PC_I \cdot PC_{I'} + c_1}{PC_I^2 + PC_{I'}^2 + c_1}, \tag{7}$$

$$S_{GM} = \frac{2GM_I \cdot GM_{I'} + c_2}{GM_I^2 + GM_{I'}^2 + c_2}. \tag{8}$$

Constants c_* are used for computation stability.

3.4 Visual information fidelity: VIF

Based on the information theory, Sheikh et al. proposed a visual fidelity metric for image quality assessment [11]. It is seen as a measurement used to quantify the level of information that can be extracted by the brain from a given scene. This metric thus relies on natural scene statistics (NSS), HVS properties and a distortion model. This metric comes from the extension of a previous work by the same authors [37] in which they proposed an information theoretic criterion for image fidelity based on NSS.

The assumption behind the VIF metric is that the random field (RF) from a wavelet decomposition subband of an image, $RF_{I'}$, can be defined as:

$$RF_{I'} = G.RF_I + V \tag{9}$$

where RF_I is the random field of the subband from the reference image, G is a deterministic scale gain field, and V is a stationary additive zero-mean Gaussian noise random field.

We have chosen to shorten the description of this metric because its mathematic demonstration is long and is difficult to summarize. The reader can refer to papers [11], [37] for a complete description and demonstration.

3.5 Information weighted SSIM: IW-SSIM

Several works have been proposed to improve computational metrics such as MSE, PSNR or SSIM by adding a weighting/pooling stage that makes behavior closer to human judgement. Information weighted SSIM introduced by Wang et al. [38] is one of them and is based on the information theory under the assumption that image components containing more information would attract more visual attention. Based on the multi-scale extension of SSIM including the weighting factors β_*, it can be expressed as follows:

$$IW\text{-}SSIM = \prod_{j=1}^{M} (IW\text{-}SSIM_j)^{\beta_j}. \tag{10}$$

By considering $I_{j,i}$ and $I'_{j,i}$ the i^{th} local image patches at the j^{th} scale, and N_j the number of windows at a given scale, the j^{th} scale IW-SSIM is defined by (see section 3.2 for the definition of the different features),

$$IW\text{-}SSIM_j = \frac{\sum_i w_{j,i} c(I_{j,i}, I'_{j,i}) s(I_{j,i}, I'_{j,i})}{\sum_i w_{j,i}}, \tag{11}$$

when $j < M$ and,

$$IW\text{-}SSIM_j = \frac{1}{N_j} \sum_i l(I_{j,i}, I'_{j,i}) c(I_{j,i}, I'_{j,i}) s(I_{j,i}, I'_{j,i}), \tag{12}$$

when $j = M$. $w_{j,i}$ is the information content weight obtained at the i^{th} spatial location in the j^{th} scale. These weights are derived by modeling the distortion channel, the perceptual channel and by considering the mutual information between the images. For more information on how to derive the aforementioned weights from the information theory, the reader is referred to the original paper [38].

3.6 Visual difference predictor (VDP): HDR-VDP-2

The VDP metric introduced by Scott Daly [8] using an interesting HVS simulation, is only applicable to low-dynamic range (LDR) images and its complexity is a real issue. An extension of VDP to higher dynamic ranges, called HDR-VDP, has been proposed by Mantiuk et al. [39], [40]. While it operates on the full range of luminance, it cannot be applied to strongly distorted images, since it is considered as a near-threshold metric.

Despite its name, HDR-VDP-2 is considered to be a breakthrough solution in comparison to the aforementioned metrics [12]. It provides three types of maps and relies on both a comprehensive model of HVS characteristics and a sound extension to a broad range of viewing conditions. Hence, the visual difference predictor models the optical and retinal pathway taking into account 1) the light scattering occurring at various levels, especially with HDR scenes, and 2) the spectral sensitivity of rods and LMS cones in addition to the luminance masking effect due to their regulation of the incoming light. At a higher level of the HVS, HDR-VDP-2 considers the overall noise affecting each subband of the multi-scale decomposition as an accumulation of: 1) a signal independent noise obtained from the contrast sensitivity function measurements, and 2) a signal dependent noise related to contrast masking.

In this paper, we used the output Q described by equation 13 derived from the pooling strategy proposed by the authors.

$$Q = \frac{1}{F.O} \sum_{f=1}^{F} \sum_{o=1}^{O} w_f \; log \left(\frac{1}{I} \sum_{i=1}^{I} D^2[f,o](i) + \epsilon \right), \quad (13)$$

where i is the pixel index and I is their total number. $D[f,o]$ is the noise-normalized signal difference for the f^{th} spatial frequency band and o^{th} orientation, w_f is the per-band weighting and $\epsilon = 10^{-5}$ is used to avoid computation instabilities.

4 Parameters of the study

In the image-based comparison scenario, the resulting correlation with mean opinion scores (MOS) (i.e. the visual fidelity prediction performance) may be influenced by several parameters, such as the rendering algorithm used, the number of 2D views or the lighting conditions. The goal of the present study is to evaluate the impact of these parameters on the performance of the image-based metric and to provide the set of values achieving the highest possible correlation with the MOS over all the available databases. Our study considers a set of cameras uniformly distributed around the object with different ways of combining their results, different positions of light sources and different means of computing mesh normals for rendering. The following subsections present details on these parameters.

4.1 Generation of 2D views and lighting conditions

In all our experiments we used 42 cameras placed uniformly around the object using a one-level dyadic split of a regular icosahedron. Each view contained a single white directional light source, fixed with respect to the camera. The used light directions were front ($[0,0,-1]$), top ($[0,-1,0]$) and top-right ($[1,-1,-1]$). Apart from these three possibilities, we also evaluated the performance of the metrics using all three conditions together (i.e we consider the mean of quality scores from the three light directions). Figure 2 illustrates some lighting conditions.

4.2 Image rendering protocols

Since a particular choice of a rendering protocol may emphasize some particular artifacts, it seems reasonable to expect that the best results will be obtained by replicating the rendering scenario used to gather the subjective data. On the other hand, the *intrinsic* visual quality of a mesh is intuitively independent from the used rendering parameters. In most of the databases of subjective experiments, similar conditions were used to render the images that were presented to the users, although there were some minor differences. Generally, we follow the common scenario used to gather subjective data. We have thus chosen to use a Phong-like local illumination shading model, with no specular and ambient reflection. We render the objects with uniform white surface reflectivity. Our rendering algorithm does not compute shadows in any form, and does not involve any kind of global illumination or texture. The resolution of the rendered images is 1024×1024.

None of the subjective experiments in the databases used normals that were explicitly stored, normals were always computed from the meshes. Some of the experiments used the standard DirectX routine for normal computation (referred to as n_0 below). In our experiments, this method has produced some small rendering artifacts, especially in the vicinity of small and elongated triangles. Other possibility is to compute the normals by averaging the normals of triangles incident each vertex (method n_1). While this approach produced artifact-free results, it is questionable whether or not it justifies applying it instead of the method used in the user studies. We have therefore decided to perform the experiments with both algorithms. Figure 2 illustrates the effects of these two normal computation methods.

These rendering parameters may seem naive respect to modern rendering pipelines. However, when evaluating the quality of a 3D model, one generally does not know the future complex shading or lighting which will be used in the real use case of this model. Hence, the most convenient approach for estimating this visual quality using an image-metric is to consider the simplest possible rendering.

Fig. 2. Different rendered views of the 3D Bimba model with noise (object from the Masking Database [41]). From left to right: *front* lighting and n_0 normal computation, *front* lighting and n_1 normal computation, *top-right* lighting and n_1 normal computation, *top* lighting and n_1 normal computation

4.3 Pooling algorithms

For a given pair of 3D objects to compare, we apply 2D metrics on the 42 pairs of corresponding views (for a given lighting condition and normal computation parameter. However since we have to derive one single visual distortion score from this set of 2D measurements, we need some pooling algorithms. We propose using the *Minkowski* norm, very popular for spatial pooling in quality metrics, defined as follows:

$$d^{mod} = \left(\frac{1}{42} \sum_{i=1}^{42} \left| d_i^{im} \right|^p \right)^{\frac{1}{p}} \quad (14)$$

where d^{mod} is the final 3D distortion score for the pair of 3D models and d_i^{im} the distortion score obtained for the i^{th} pair of images. In our study we consider $p = 1$, $p = 2$, $p = 3$ and $p = \infty$ (i.e. we consider the maximum image distortion score).
The views that we consider for applying the 2D metrics are regularly sampled over a bounding sphere. However some views may be far more visually important than others and thus should have more weight on the distortion/fidelity prediction. To test this hypothesis we also considered a weighted pooling (noted as p_w in the experiment section):

$$d^{mod} = \frac{\sum_{i=1}^{42} w_i d_i^{im}}{\sum_{i=1}^{42} w_i} \quad (15)$$

where w_i is the weight assigned to the i^{th} pair of images. To compute these weights, we rely on a recent work on automatic viewpoint preference selection [42]. The authors determine, using learning, a set of attributes highly relevant for predicting the best viewpoint of a given 3D model. We use the most efficient attribute, which is surface visibility, introduced by Plemenos and Benayada [43]. It refers to the ratio of visible surface in the view, to the total surface of the 3D object. In practice, to compare a distorted model with a reference one, we compute these weights on the views of the reference model. Figure 3 illustrates these weights for different views of the Bimba 3D model.

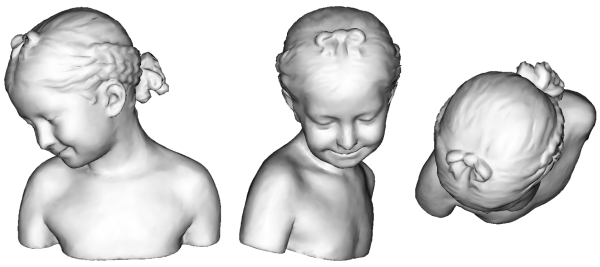

Fig. 3. Different rendered views of the 3D Bimba model with noise (object from the Masking Database [41]). From left to right, viewpoint importance weights are equal to 1, 0.65 and 0.47.

5 Mean Opinion Score Databases

In order to assess and compare the performance of quality assessment metrics (image-based and model-based), we compute correlation of metric predictions with average human opinion. Human opinion is represented by mean opinion scores (MOS) obtained through subjective rating experiments. To the best of our knowledge, there are currently four publicly-available subject-rated 3D model databases, all of which we consider in our comparisons. These databases, detailed below, contain different categories of distorted and reference objects, as well as the corresponding mean opinion scores reflecting their qualities collected from human subjects.

The LIRIS/EPFL General-Purpose Database [44] was created at the EPFL, Switzerland. It contains 88 models with between 40K and 50K vertices generated from 4 reference objects (Armadillo, Dyno, Venus and RockerArm). Two types of distortion (noise addition and smoothing) are applied with different strengths and at four locations: uniformly (on the whole object), on smooth areas, on rough areas and on intermediate areas. These distortions aim at simulating the visual impairment of generic geometric processing operations (compression, watermarking, filtering). 12 observers participated in the subjective evaluation; they were asked to provide a score reflecting the degree of perceived distortion between 0 (identical to the original) and 10 (worst case).

The LIRIS Masking Database [41] was created at the Université of Lyon, France. It contains 26 models with between 9K and 40K vertices generated from 4 reference objects (Armadillo, Bimba, Dyno and Lion) specifically chosen because they contain large smooth and rough areas. The only distortion is random noise addition, applied at three strengths. However, it is applied either on smooth or rough regions. The specific goal of this database was to evaluate the *visual masking* effect. It turns out that the noise is indeed far less visible on rough regions. Hence, metrics should follow this perceptual mechanism. 11 observers participated in the subjective evaluation.

The IEETA Simplification Database [45] was created at the University of Aveiro, Portugal. It contains 30 models generated from 5 reference objects (Bunny, Foot, Head, Lung and Strange), with a complexity ranging from 2K to 25K vertices. The reference models have been simplified using three different methods at two levels (20% and 50% of the original number of faces). 65 observers participated in the subjective evaluation; they were asked to provide a score from 1 (very bad) to 5 (very good).

The UWB Compression Database [15] was created at the University of West Bohemia, Czech Republic. It contains 68 models created from 5 reference meshes. For each reference mesh, all the model versions share the connectivity of the original. The main purpose of the database is to evaluate different kinds of artifacts introduced by different compression algorithms. In contrast to previous experiments, instead of scoring, a binary forced choice paradigm has been adopted when collecting the user opinions. This meant that for each of the 69 users in the test, triplets of meshes were presented, with one mesh designated as original, and two randomly chosen distorted versions. The users were asked to select the preferred version of the two distorted ones. The collected data are available both in terms of user choices and in

scores computed from the choices in a manner described in [15].

6 Results and evaluation

In this section, we first provide a statistical and quantitative analysis of the effect of the different parameters involved in image-based quality assessment. This study provides useful recommendations concerning the best parameters to use in order to obtain *optimal* image-based metrics. We then provide a comprehensive comparison of these optimal image-based metrics with the best state-of-the-art model-based metrics.

6.1 Analysis of the parameters

After presenting the methodology of our statistical analysis, the following subsections evaluate the impact of each image generation parameter on the performance of image-based quality assessment.

6.1.1 Data and methodology

Our study shares many similarities with full factorial experiments [46] (also called fully crossed experiments) considerably used in statistics to study the effect of different factors on a response variable. In this case we consider four factors: the metric (7 possible values), the lighting (4 possible values), the pooling (5 possible values) and the rendering (2 possible values). For each of the $7 \times 4 \times 5 \times 2 = 280$ possible combinations, we obtain one metric and thus one visual fidelity value for a given pair of 3D models compared. Our goal is to find, for each factor, the values providing the best results (i.e. the best metrics, lighting conditions, etc). In practice, for each of the 280 possible combinations of parameters, we compute the Spearman's rank order correlation between the obtained fidelity scores and the mean opinion scores over the objects of each of the four databases presented in the previous section. For each database, we consider two response variables that correspond to two scenarios:

1) The Spearman correlation computed over the whole set of 3D models. Note that the compression database is not considered here because the data acquisition procedure used to obtain this database does not capture inter-model relations.
2) The Spearman correlation is computed separately for each class of objects (a class of objects is a set of distorted versions of a single original). The correlations are then averaged. This corresponds to an easier scenario for the metrics, since they do not have to rank different objects with different distortions, but only evaluate different distortions applied to a single object.

We thus have three Spearman coefficients for each parameter combination in the case of scenario-1 (one per database, the compression database is not suitable for this scenario), and four in the case of scenario-2. We consider the Spearman correlation as our performance indicator since it is widely used to assess the performance of quality assessment metrics for image, video and graphics.

In practice, for a given factor associated with n possible values, we have n sets of $\frac{280\times3}{n}$ (in the case of scenario-1) or $\frac{280\times4}{n}$ (in the case of scenario-2) paired Spearman coefficients that we can analyze and compare. These sets of coefficients are paired, since they correspond to the same combinations of the remaining factor's values. To estimate the effect of a given factor on objective metric performance, we conduct pairwise comparisons of each of its values with the others (i.e. $\frac{n(n-1)}{2}$ comparisons). Since the corresponding sets of Spearman coefficients are paired, we can conduct a more thorough analysis than simple mean or median comparison. To assess the superiority of some factor values among others, we first consider a statistical significance test. Usually, a paired Student's t-test is used for this kind of analysis, however our data are not physical or experimental measurements but correlations and thus do not satisfy the normality assumption. Hence, we consider instead the non-parametric equivalent which is the Wilcoxon signed-rank test [47].To obtain more quantitative information, we also compute the median of paired differences, as well as the 25th and 75th percentiles. These measurements are used to quantify the relative efficiency of each factor value against the rest. Results are presented in matrix form in the tables below. Medians and percentiles are multiplied by 100 to improve legibility. Factor values are ranked in the first column regarding their superiority to others (i.e. each ranked metric has a positive median of differences with all metrics below). For each factor value presented at the left of each row, the medians of paired differences against the others are listed in the rest of the row. Cases for which the difference is statistically significant with p-value below 0.05 are highlighted in gray. The following sections present the results of this analysis for each factor.

6.1.2 Influence of the metrics

Tables 1 and 2 detail the median of Spearman correlation differences for each value pair of the metric factor. It is interesting to observe that the ranking remains similar in both scenarios. In the more difficult one (presented in table 1), IW-SSIM outperforms all the other metrics at a statistically-significant level; in particular, it shows a median improvement of Spearman correlation of 0.085 and 0.095 compared to MS-SSIM and FSIM which are ranked just behind. When correlations are computed separately for each class of objects (see table 2) the differences between metrics are attenuated. These three metrics (IW-SSIM, MS-SSIM and FSIM) are significantly better than MSE and PSNR with more than 0.10 median Spearman improvement in both scenarios. The cases of VIF and HDR-VDP-2 form a particular case since these metrics produced very heterogeneous results, as illustrated by

the percentiles, depending on the considered database. In practice, the VIF metric yielded very good results for the Masking database but poor results on the others. On the contrary, HDR-VDP-2 outperformed its counterparts for the Compression database but produced the worst results for the Masking one.

The ranking that we obtain is coherent with recent studies that evaluate the performance of 2D metrics on databases of natural images [30], [36]; in these evaluations, the best results are obtained by IW-SSIM, FSIM and MS-SSIM, which are significantly better than VIF, and also significantly better than PSNR. The HDR-VDP-2 metric was shown to outperform MS-SSIM for natural images in [12]. However, this metric is based on a complex visual model of early human vision and seems to be not adapted to the specificity of some geometric artifacts. In particular, we noted a high sensitivity to how the normals are computed.

6.1.3 Influence of lighting

Table 3 presents the median of correlation differences for the lighting factor, for correlations computed over whole corpuses. In contrast with the metrics, the difference in performance between the different parameters is much less. However, we can still draw some interesting conclusions. Indirect illuminations (*Top* and *Top-Right*) provide significantly better results than direct ones (*Front*). This appears reasonable, since it emphasizes artifact visibility, a fact also observed in the study conducted by Rogowitz and Rushmeier [5]. If we look at the median difference of the Spearman correlation, the *Top* lighting condition is better than the *Top-Right* one, however the median of correlation differences is very low (0.007) and the Wilcoxon test is not significant (p-value=0.77). Besides the fact that indirect lighting facilitates the task of the metrics by emphasizing the artifacts, this superiority of the *Top* lighting condition could be linked to previous perception results [48], [49] demonstrating that the visual system assumes light is above when viewing the image of a shaded 3D surface. Hence people could be more sensitive to artifacts emphasized by lighting from the top.

We do not include the table for correlations computed and averaged per class of models, since in this case the lighting conditions influence the results even less and no significant difference is observed.

TABLE 3
Median of differences between Spearman correlations for the different pairs of lighting conditions. Correlations are computed over the whole set of models for each database.

	Top-Right	Mean	Front
1 Top	**+0.7** [-1.3;2.4]	**+0.6** [-1.1;2.2]	**+1.1** [-1.5;7.2]
2 Top-Right		**+0.1** [-0.9;1.3]	**+1.3** [-0.7;4.3]
3 Mean			**+0.7** [-0.7;4.7]
4 Front			

6.1.4 Influence of pooling

Table 4 presents the median of correlation differences for the pooling factor, when correlations are computed and averaged per class of models. The main conclusion that can be drawn is that, for this setting, p_w, p_1, p_2 and p_3 are significantly better than p_∞. There is no statistically significant difference between p_{1w}, p_1, p_2 and p_3. In particular, the weighted pooling does not improve the results. Note that we conducted tests by taking w_i^2 as weights in order to enhance their influence, however the results were similar.

When we consider the correlations computed over the whole set of models, then no significant differences are observed between the different conditions (p-values are all above 0.45). It is actually difficult to determine a best pooling parameter in an absolute way because this best parameter greatly depends on the metrics and their respective sensitivity. For IW-SSIM, MS-SSIM and FSIM the best setting is p_∞. On the contrary, this setting produces the worst setting for HDR-VDP-2, VIF, MSE and PSNR.

The previous study by Cleju and Saupe [7] about image-based 3D model quality evaluation only considered the p_1 case.

TABLE 4
Median of differences between Spearman correlations for the different pairs of pooling conditions. Correlations are computed and averaged per class of models for each database.

	p_{1w}	p_2	p_3	p_∞
1 p_1	**+0** [-0.1;0.5]	**+0** [0;0.3]	**+0** [0;0.5]	**+1.7** [-0.4;4.9]
2 p_{1w}		**+0** [-0.3;0.3]	**+0** [-0.4;0.3]	**+1.5** [0;4.9]
3 p_2			**+0** [0;0.1]	**+1.8** [-0.3;4.6]
4 p_3				**+1.4** [-0.2;4.6]
5 p_∞				

6.1.5 Influence of normal computation

For normal computation, the best parameter is n_0 (i.e. DirectX routine), with a median of correlation differences equal to +1.4 [-0.8; 3.3] (values are multiplied by 100 as in the tables). These values are obtained when correlations are computed over the whole corpuses. If we consider per-class correlations, the rank is the same but the difference is less: +0.7 [-0.3; 2.9]. Just as for lighting, this relative superiority is explained by the fact that the n_0 condition emphasizes artifact visibility. However for these two scenarios, the superiority of n_0 does not reach statistically significant levels (p-values resp. equal to 0.69 and 0.73). Just as for pooling, the main reason for these high p-values is that the best parameter greatly depends on the metric: all metrics, except HDR-VDP-2, are better with the n_0 parameter which emphasizes artifact visibility. However HDR-VDP-2 is too sensitive to evident artifacts due to normal flips (as illustrated in figure 2, left) and thus yields much better results when normals are slightly smoothed (n_1 parameter).

TABLE 1
Median of differences between Spearman correlations for the different pairs of metrics (25th and 75th percentiles are also presented in brackets). Correlations are computed over the whole set of models for each database. For readability, medians and percentiles are multiplied by 100. We highlight in gray cases for which the difference reach a statistically-significant level.

	MSSIM	FSIM	VIF	HDR-VDP-2	MSE	PSNR
1 IW-SSIM	**+8.5** [-0.9;11.3]	**+9.5** [0.9;13.9]	**+3.5** [-7.9;21.5]	**+12.4** [3.8;24.1]	**+11.3** [5.1;31.6]	**+12.2** [6.4;31.0]
2 MS-SSIM		**+2.5** [-1.2;5.3]	**+7.0** [-21.3;13.9]	**+9.6** [-2.4;17.4]	**+7.5** [2.2;23.0]	**+10.3** [1.0;22.7]
3 FSIM			**+5.4** [-18.1;8.6]	**+8.7** [-4.4;17.3]	**+9.3** [2.2;18.0]	**+11.3** [3.3;17.6]
4 VIF				**+6.5** [-11.6;32.9]	**+10.4** [1.6;15.9]	**+10.1** [3.8;15.6]
5 HDR-VDP-2					**+0.2** [-13.2;21.9]	**+2.1** [-12.0;21.6]
6 MSE						**+0.0** [-0.5;0.8]
7 PSNR						

TABLE 2
Median of differences between Spearman correlations for the different pairs of metrics. Correlations are computed and averaged per class of models for each database.

	MSSIM	FSIM	VIF	HDR-VDP-2	MSE	PSNR
1 IW-SSIM	**+0.0** [-1.1;2.4]	**+1.1** [-1.4;7.9]	**+6.9** [-2.2;14.9]	**+2.4**[-6.9;24.0]	**+8.6** [3.0;19.2]	**+10.7** [2.3;20.3]
2 MS-SSIM		**+1.1** [0.0;4.1]	**+6.9** [-2.0;12.6]	**+1.1** [-6.6;22.3]	**+11.0** [1.4;16.9]	**+11.5**[1.4;17.8]
3 FSIM			**+5.7** [-4.3;8.4]	**+0.2** [-9.8;19.7]	**+9.2** [1.1;14.3]	**+10.4** [-1.1;14.4]
4 VIF				**-8.0** [-16.7;20.6]	**+2.6** [-4.6;10.8]	**+3.6** [-3.4;14.4]
5 HDR-VDP-2					**+10.0** [-17.4;24.2]	**+9.7** [-17.1;25.4]
6 MSE						**+0.0** [0.0;1.5]
7 PSNR						

6.1.6 Recommendations

The results presented above draw some conclusions regarding the image-based evaluation of 3D visual fidelity:

- IW-SSIM yields the best results. Simple measurements like PSNR and MSE should be avoided.
- VIF and HDR-VDP-2 may produce excellent results depending on the database.
- For lighting, an indirect illumination (top or top-right) should be preferred since frontal lighting tends to mask the artifacts.
- p_∞ pooling yields the worst results for MSE, HDR-VDP-2 and VIF and the best ones for IW-SSIM, MS-SSIM and FSIM. Viewpoint weights bring no gain.
- DirectX normal computation yields better results than simple face normal averaging, since it also tends to emphasize the visual artifacts. However, HDR-VDP-2 prefers the second method since it removes some rendering artifacts due to degenerate triangles.

6.2 Comparison with model-based metrics

Now that we have analyzed the influence of the different image generation parameters for the task of image-based quality assessment of 3D graphical objects, our goal is to compare the performance of this framework with the best existing model-based metrics.

6.2.1 Methodology

To measure the performance of the metrics, we consider the correlations with mean opinion scores for the four databases presented in section 5 and used above. We consider the Spearman Rank Order Correlation Coefficient (SROCC), which measures the monotonic association between the MOS and the metric values and the Pearson Linear Correlation Coefficient (LCC), which measures prediction accuracy. The Pearson correlation is computed after performing a non-linear regression on the metric values as suggested by the video quality experts group (VQEG) [50], using a cumulative Gaussian function. This serves to optimize the matching between the values given by the objective metric and the subjective opinion scores provided by the human subjects. This step allows the evaluation to take into account the saturation effects typical for human senses.

Just as when analyzing image-based parameters, we consider two scenarios: either the correlation is computed by considering all classes of objects from the corpus together, or by considering one class at a time and then averaging the results. In the latter case, the metrics produce better results, since they only have to evaluate and predict distortions applied on the same reference object. Both scenarios represent realistic use-cases. Results for these two scenarios are presented in tables 5 and 6. For image-based metrics we present two correlation values per column: the main one is obtained by learning the parameters (pooling, lighting, rendering) using a n-folds cross validation on the databases ($n = 3$ or $n = 4$ depending on the scenario), while the values in brackets are computed for the best parameter configuration. The n-folds cross validation is performed as follows: for each metric and each database, we consider the parameters that

TABLE 5
Performance comparison of model-based (top) and image-based (bottom) metrics on 3 benchmark databases. For each database, correlations are computed over all models.

	Masking		Simplification		General Purpose	
	SROCC	LCC	SROCC	LCC	SROCC	LCC
RMS	48.8	41.2	64.3	58.6	26.8	28.1
MSDM2 [17]	89.6	87.3	86.7	89.2	80.4	81.4
FMPD [20]	80.2	80.8	87.2	89.3	81.9	83.5
DAME [15]	68.1	58.6	NA	NA	76.6	75.2
FSIM [36]	53,2(61)	45,5(54,6)	64,7(79,8)	67,5(76,9)	50,4(55,1)	58(63,3)
IW-SSIM [38]	66,6(68,3)	64,5(67,2)	64,4(82,8)	70,5(79,8)	67,5(69,8)	70,1(73,6)
MS-SSIM [35]	41,6(56,7)	34,8(48,6)	66,5(84,4)	67,9(79,3)	56,1(61)	62,5(68,2)
VIF [11]	75,7(77,1)	76,4(76,7)	64,2(71,2)	62,4(72,1)	45,2(45,9)	49(50,1)
HDR-VDP-2 [12]	32,4(50,1)	26,6(52,7)	60,4(63,9)	66,6(67)	69,6(71,7)	66,3(67,7)
MSE	58,1(62,1)	17,9(36)	61,5(67,1)	61,7(66,3)	34,4(36,8)	38,8(41,2)

TABLE 6
Performance comparison of model-based (top) and image-based (bottom) metrics on 4 benchmark databases. For each database, correlations are computed per class of models.

	Masking		Simplification		General Purpose		Compression	
	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC
RMS	70	58.3	80.6	80.4	40.1	40.5	52	49
MSDM2 [17]	95.7	91.8	85.1	95.4	86.6	86.4	78	89.3
FMPD [20]	94.3	95.9	80.6	91.2	85.3	85.2	81.8	88.8
DAME [15]	93.7	95.7	NA	NA	82.3	82.3	85.6	93.5
FSIM [36]	85,7(95,7)	85,5(94,5)	92(92)	92(92,5)	62,1(64)	64,4(64,7)	69,1(73,2)	75,2(76,4)
IW-SSIM [38]	82,9(90)	80,6(89,5)	90,9(94,3)	93,4(94,6)	68,9(74)	72,2(77,5)	65,5(68,9)	73,6(74,2)
MS-SSIM [35]	85,7(90)	79,2(87,9)	90,9(94,3)	93,4(93,7)	67(68,6)	68,5(70,1)	68,5(72,4)	76,6(77,1)
VIF [11]	92,9(92,9)	89,3(90,6)	77,1(87,4)	92(95,8)	52,7(55,6)	57,3(59,6)	53(64,6)	61,7(74,2)
HDR-VDP-2 [12]	55,7(84,3)	59,4(77,9)	89,7(89,7)	93,7(97,4)	77,3(77,3)	73,5(73,5)	74,9(79,3)	78,7(84,3)
MSE	68,6(91,4)	67,2(86,9)	85,1(90,9)	85,4(90,9)	45,8(51,8)	48,7(56,1)	45,7(61,4)	48(64,7)

maximize its performance on the $n-1$ other databases. We present the results for the best model-based metrics (MSDM2, FMPD and DAME) and best image-based metrics (IW-SSIM, MS-SSIM, FSIM, VIF and HDR-VDP-2). We also include results of RMS (model-based) and MSE (image-based) as baselines. The best performing image-based and model-based metrics are highlighted for each database.

6.2.2 Results

The tables confirm the superiority of IW-SSIM; this metric is not always ranked first but produces remarkably stable results (at least 0.65 correlation whatever the database or the scenario). HDR-VDP-2 also produces excellent results that outperform other image-based metrics on the General-Purpose and Compression databases, when correlations are computed per class of models. However its performance is very poor on the Masking database. An opposite behavior is observed for the VIF metric (excellent on the Masking database and very poor on others).

When comparing the results of image-based and model-based metrics, we observe that in simple scenarios (per-class correlation and one single type of distortion), best image-based metrics perform very well regarding model-based ones (correlations around 90-95%, see table 6, Masking and Simplification columns). However, for difficult scenarios, such as evaluating different kinds of distortions (such as general-purpose and compression databases) and/or different 3D models together (as presented in table 5), recent model-based metrics are significantly better. The best image-based metrics provide a correlation of around 65-75% as opposed to 85-90% for the best model-based ones. These results show that image-based metrics are excellent for ranking different versions of a single object under a single type of distortion, however, they are less accurate at differentiating the artifact visibility between different distortions, or distortions applied on different 3D models.

6.2.3 Similarities and differences with previous studies

Our results confirm the findings from studies from Watson et al. [6] and Cleju and Saupe [7] showing that, for simplification artifacts, the image-based MSE performs better than the model-based RMS. This effect appears clearly in our results when considering the best set of parameters for MSE; for the simplification database, we obtain a Pearson correlation of 0.91 (resp. 0.67) for MSE against 0.80 (resp. 0.59) for RMS when correlation is computed per class (resp. over all models). In addition to this confirmation, our results demonstrate that this superiority of the image-based MSE over the model-based RMS is also true for other types of distortions (see the results for the other databases). However the MSE is very sensitive to the parameters used for generating

images as can be seen by the great difference which may appear between correlation values for the best parameters and the cross-validation ones.
In the experiments from Watson et al. [6] and Cleju and Saupe [7], perceptually-based metrics like the Bolin and Meyer model [23] and SSIM [10] did not demonstrate any clear superiority over MSE. When looking at the Simplification database, our results remain quite close to these observations: the MS-SSIM (multi-scale version of SSIM) demonstrates a moderate improvement over MSE while the HDR-VDP-2 (the most closely related to Bolin and Meyer's model) remains quite close to MSE. Interestingly, our results show that when evaluating different kinds of distortions (such as for general-purpose and compression databases) then the performance of MSE drops significantly, while perceptual metrics (including MS-SSIM and HDR-VDP-2) still remain good. As an illustration, for the Compression database, the Pearson correlation for MS-SSIM and HDR-VDP-2 is around 0.78 as opposed to 0.48 for MSE.

7 Benefits and applications of image-based metrics

We have shown in the previous section that several image metrics (e.g. IW-SSIM, MS-SSIM, HDR-VDP-2) may be excellent predictors of the visual quality of 3D models, depending on the nature of the artifacts. In spite of the simplicity of the rendering used for generating images, the performance achieved by these metrics is comparable with model-based metrics under certain conditions. That is actually a quite remarkable finding since a user who has better information on the character of the final rendering may achieve even better results. In this section we detail the benefits of the image-based approach as well as several applications.

7.1 Benefits

In addition to their good performance, image-based metrics offer several significant benefits over model-based ones:

- They are able to handle non-manifold meshes as well as maps used for simulating geometric details such as normal maps. They can even evaluate distortions on other representations than triangular meshes (e.g. NURBS or implicit surfaces).
- They can compute view-direction specific fidelity measurements, which may be useful when the considered object will only be viewed from a particular direction (e.g. objects standing on the ground in a virtual scene).
- As stated above, any information about the future rendering pipeline of the 3D model (lighting, material) is easy to integrate to generate the views and then improve measurement accuracy. Other types of information can be included easily in the same framework: texture, colors, etc. without the need for complex metric combinations.

7.2 Applications

We present below two applications that would be impossible to conduct with model-based metrics.

7.2.1 Comparison of simplification algorithms for non-manifold meshes

In this first scenario, the goal is to compare the performance of two edge-collapse simplification algorithms, applied on a *Tree* model created using a modeling software: the first is based on the Quadric Error Metric (QEM) [51], while the second is based on Local Hausdorff distances (LH) [52]. Model-based metrics (except RMS) cannot be applied for evaluating such non-manifold models, mostly because they rely on differential geometry operators. IW-SSIM has demonstrated excellent performance for evaluating simplification artifacts (significantly better than RMS). Hence, we consider this image metric for this task (with n_0 normal calculation, *top* lighting and p_1 pooling as recommended by our study). Results, illustrated in figure 4, show that in the simplification range between 20% and 80%, the LH method [52] performs better than QEM [51].

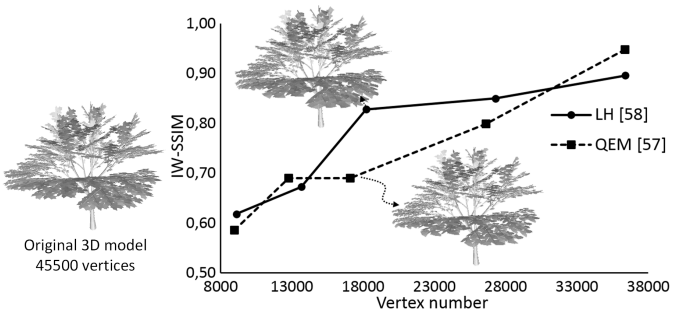

Fig. 4. Simplification rate vs visual fidelity for two simplification algorithms applied on the non manifold *Tree* model. Fidelity is measured using the IW-SSIM image metric (higher is better).

7.2.2 Evaluation of the visual impact of normal map compression

In this application, we consider scanned 3D models defined as manifold meshes along with normal maps. The need to decrease the data size (for remote Web-based visualization for instance) requires compression of the normal map which is often larger in size than the geometry. However, a model-based metric is not able to evaluate the impact of normal map compression on model appearance. In the same way as for the previous application, we use the IW-SSIM metric for this task. However, this time, we consider that the material and lighting environment of the final rendering pipeline and the preferred viewpoint are known. The curve in figure 5 illustrates the JPEG compressed size of the normal map versus the measured visual fidelity for the Squirrel model from the EPFL Computer Graphics and Geometry Laboratory. The original sizes of the geometry and

normal map are, respectively, 230KB (binary PLY) and 2800KB (visually lossless JPEG i.e. Q=100). The figure shows a quick drop in quality when allocating less than 200-400KB for the normal map, suggesting that such sizes would lead to a good tradeoff between quality and transmission time.

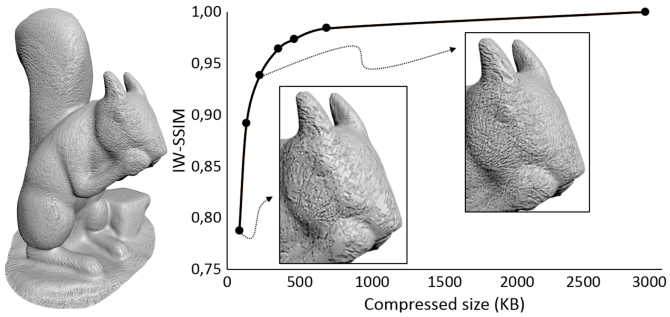

Fig. 5. Compressed size vs visual fidelity when compressing the normal map of the *Squirrel* model (6K vertices). Fidelity is measured using the IW-SSIM image metric.

8 Conclusion

In this paper, we investigated the use of image metrics for assessing the visual fidelity of 3D graphical objects. First, we determined the best parameters for this kind of evaluation (among different types of lighting conditions, pooling methods, normal computation) and the most efficient metrics. We then compared this image-based approach to the best performing model-based metrics for quality assessment of 3D objects, in terms of correlation with human judgment.

Our results suggest that, despite the very simple rendering scenario used in the experiment, image-based metrics perform very well in evaluating the quality of different versions of a same object under a single type of distortion. In such scenarios, they come remarkably close to the performance of model-based distortion measurements. However, they are less accurate in comparing different distortions or distortions applied to different 3D models. Hence, for simple use cases (e.g. determining the best parameters of a compression algorithm), or in cases when model-based metrics cannot be applied (non-manifold meshes, meshes with normal maps) image-based metrics will work very well. However, in scenarios involving the ranking of different distortions applied on different 3D objects (e.g. benchmarking different watermarking algorithms run on different sets of models) the model-based approaches perform better. These metrics, however, cannot include other rendering effects, such as textures or surface shaders, in such a straightforward way as image-based metrics do. We finally illustrated several applications for which image-based approaches are particularly well-suited.

The present study concerns visual fidelity of 3D geometric objects without texture or complex rendering attributes (e.g. light fields). A similar study could be conducted for such complex data. However, our assumption is that our findings should remain valid; i.e. image-based approaches will certainly have the same benefits/drawbacks as in our study. Many recent works concern the quality assessment of images created by graphic rendering [53], [54], [55]. Their goal is to detect the artifacts introduced by the rendering pipeline (e.g. structured noise from approximate global illumination). In the future, we plan to investigate new kinds of metrics taking into account both image rendering artifacts and model artifacts (i.e. distortion on geometry or texture) in order to produce global fidelity/quality indices.

Acknowledgments

This work was supported by the European Regional Development Fund (ERDF), project "NTIS New Technologies for the Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

References

[1] G. Lavoué and M. Corsini, "A comparison of perceptually-based metrics for objective evaluation of geometry processing," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 636–649, 2010.

[2] M. Corsini, M. C. Larabi, G. Lavoué, O. Petřík, L. Váša, and K. Wang, "Perceptual Metrics for Static and Dynamic Triangle Meshes," *Computer Graphics Forum*, vol. 32, no. 1, pp. 101–125, 2013.

[3] P. Lindstrom and G. Turk, "Image Driven Simplification," *ACM Transactions on Graphics*, vol. 19, no. 3, pp. 204–241, 2000.

[4] Z. Wang and A. Bovik, *Modern image quality assessment*. Morgan & Claypool Publishers, 2006, vol. 2, no. 1.

[5] B. E. Rogowitz and Holly E. Rushmeier, "Are image quality metrics adequate to evaluate the quality of geometric objects?" *Proceedings of SPIE*, pp. 340–348, 2001.

[6] B. Watson, A. Friedman, and A. McGaffey, "Measuring and predicting visual fidelity," *ACM Siggraph*, pp. 213–220, 2001.

[7] I. Cleju and D. Saupe, "Evaluation of supra-threshold perceptual metrics for 3D models," in *Symposium on Applied Perception in Graphics and Visualization*. ACM Press, Jul. 2006.

[8] S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital images and human vision*, A. B. Watson, Ed. Cambridge: MIT Press, Oct. 1993, pp. 179–206.

[9] J. Lubin, "A visual discrimination model for imaging system design and evaluation," in *Vision Models for Target Detection and Recognition*, E. Peli, Ed. World Scient. Pub., 1995, pp. 245–283.

[10] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, p. 600?612, 2004.

[11] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[12] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "Hdr-vdp-2 : A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Siggraph*, 2011.

[13] W. Schroeder, J. A. Zarge, and W. E. Lorensen, "Decimation of triangle meshes," in *ACM Siggraph*, 1992, pp. 65 – 70.

[14] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro: Measuring Error on Simplified Surfaces," *Computer Graphics Forum*, vol. 17, no. 2, pp. 167–174, Jun. 1998.

[15] L. Váša and J. Rus, "Dihedral Angle Mesh Error: a fast perception correlated distortion measure for fixed connectivity triangle meshes," *Computer Graphics Forum*, vol. 31, no. 5, 2012.

[16] Z. Karni and C. Gotsman, "Spectral compression of mesh geometry," in *ACM Siggraph*, 2000, pp. 279–286.

[17] G. Lavoué, "A Multiscale Metric for 3D Mesh Visual Quality Assessment," *Computer Graphics Forum*, vol. 30, no. 5, pp. 1427–1437, 2011.

[18] F. Torkhani, K. Wang, and J.-m. Chassery, "A Curvature Tensor Distance for Mesh Visual Quality Assessment," in *International Conference on Computer Vision and Graphics*, 2012.

[19] M. Corsini, E. D. Gelasca, T. Ebrahimi, and M. Barni, "Watermarked 3-D Mesh Quality Assessment," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 247–256, Feb. 2007.

[20] K. Wang, F. Torkhani, and A. Montanvert, "A Fast Roughness-Based Approach to the Assessment of 3D Mesh Visual Quality," *Computers & Graphics*, 2012.

[21] D. Tian and G. AlRegib, "FQM: A Fast Quality Measure for Efficient Transmission of Textured 3D Models," in *ACM Multimedia*, 2004, pp. 684–691.

[22] Y. Pan, I. Cheng, and Anup Basu, "Quality metric for approximating subjective evaluation of 3-D objects," *IEEE Transactions on Multimedia*, vol. 7, no. 2, pp. 269–279, Apr. 2005.

[23] M. Bolin and G. Meyer, "A perceptually based adaptive sampling algorithm," in *ACM Siggraph*, 1998, pp. 299–309.

[24] L. Qu and G. Meyer, "Perceptually guided polygon reduction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 5, pp. 1015–1029, 2008.

[25] M. Reddy, "SCROOGE: Perceptually-Driven Polygon Reduction," *Computer Graphics Forum*, vol. 15, no. 4, pp. 191–203, 1996.

[26] D. Luebke and B. Hallen, "Perceptually Driven Simplification for Interactive Rendering," in *Eurographics Workshop on Rendering Techniques*, 2001, pp. 223–234.

[27] N. Menzel and M. Guthe, "Towards Perceptual Simplification of Models with Arbitrary Materials," *Computer Graphics Forum*, vol. 29, no. 7, pp. 2261–2270, Sep. 2010.

[28] Q. Zhu, J. Zhao, Z. Du, and Y. Zhang, "Quantitative analysis of discrete 3D geometrical detail levels based on perceptual metric," *Computers & Graphics*, vol. 34, no. 1, pp. 55–65, Feb. 2010.

[29] K. Seshadrinathan and A. Bovik, "Automatic prediction of perceptual quality of multimedia signals - a survey," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 163–186, 2011.

[30] L. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," *International Conference on Image Processing (ICIP)*, pp. 1477–1480, 2012.

[31] M.-C. Larabi, A. Saadane, and C. Charrier, "Quality assessment approaches," in *Digital Color*. Wiley, 2013, pp. 265–306.

[32] ——, "Quality assessment of still images," in *Advanced Color Image Processing and Analysis*. Springer New York, 2013, pp. 423–447.

[33] A. Beghdadi, M.-C. Larabi, A. Bouzerdoum, and K. Iftekharuddin, "A survey of perceptual image processing methods," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 811 – 831, 2013.

[34] Q. Huynh-Thu and M. Ghanbari, "The accuracy of psnr in predicting video quality for different video scenes and frame rates," *Telecommunication Systems*, vol. 49, no. 1, pp. 35–48, 2012.

[35] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *IEEE Asilomar Conf. on Signals, Systems, and Computers*, 2003, pp. 1398–1402.

[36] L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment." *IEEE Transactions on Image Processing*, no. 99, pp. 1–1, Jan. 2011.

[37] H. Sheikh, A. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Proc.*, vol. 14, no. 12, pp. 2117–2128, 2005.

[38] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185 – 1198, 2011.

[39] R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Visible difference predicator for high dynamic range images," *IEEE Int.Conference on Syst., Man and Cybernetics*, vol. 3, pp. 2763–2769, 2004.

[40] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel, "Predicting visible differences in high dynamic range images: model and its calibration," in *Proc. SPIE*, vol. 5666, 2005, pp. 204–214.

[41] G. Lavoué, "A local roughness measure for 3D meshes and its application to visual masking," *ACM Transactions on Applied Perception (TAP)*, vol. 5, no. 4, 2009.

[42] A. Secord, J. Lu, A. Finkelstein, and M. Singh, "Perceptual models of viewpoint preference," in *ACM Siggraph*, 2011.

[43] D. Plemenos and M. Benayada, "Intelligent display in scene modeling. new techniques to automatically compute good views," in *GraphiCon*, 1996.

[44] G. Lavoue, E. Drelie Gelasca, F. Dupont, A. Baskurt, and T. Ebrahimi, "Perceptually driven 3D distance metrics with application to watermarking," in *SPIE*, vol. 6312, Aug. 2006.

[45] S. Silva, B. Santos, and C. Ferreira, "Comparison of methods for the simplification of mesh models using quality indices and an observer study," *SPIE*, pp. 64 921L–64 921L–12, 2007.

[46] R. Fisher, *The Design of Experiments*. Edinburgh: Oliver & Boyd, 1935.

[47] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[48] J. Sun and P. Perona, "Where is the sun?" *Nature neuroscience*, pp. 183–184, 1998.

[49] J. O'Shea, M. Banks, and M. Agrawala, "The assumed light direction for perceiving shape from shading," *symposium on Applied perception in graphics and visualization*, 2008.

[50] VQEG, "Final report on the validation of objective models of video quality assessment," Tech. Rep., 2000.

[51] M. Garland and P. S. Heckbert, "Surface simplification using quadric error metrics," in *ACM Siggraph*, 1997, pp. 209–216.

[52] F. Caillaud, V. Vidal, F. Dupont, and G. Lavoue, "Progressive compression of generic surface meshes," in *Computer Graphics International*, 2015.

[53] M. Čadík, R. Herzog, R. Mantiuk, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Learning to Predict Localized Distortions in Rendered Images," in *Pacific Graphics*, vol. 32, no. 7, 2013.

[54] R. Herzog, M. Čadík, T. O. Aydin, K. I. Kim, K. Myszkowski, and H.-p. Seidel, "NoRM : No-Reference Image Quality Metric for Realistic Image Synthesis," *Computer Graphics Forum*, vol. 31, no. 2, 2012.

[55] M. Čadík, R. Herzog, and R. Mantiuk, "New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts," *ACM Siggraph*, 2012.

Guillaume Lavoué (M'11-SM'13) received his PhD from the University of Lyon (2005) where he is now an associate professor. He defended his habilitation title in 2013. His research interests include diverse aspects of geometry processing as well as quality assessment and perception for computer graphics. Guillaume Lavoué is currently chair of the IEEE SMC TC on Human Perception and Multimedia Computing and serves as an associate editor for The Visual Computer journal (Springer).

Mohamed-Chaker Larabi (M'05-SM'07) received his PhD from the University of Poitiers (2002). He is currently Associate Professor at the same university. His current scientific interests deal with quality of experience and bio-inspired processing/coding/optimization of images and videos, 2D, 3D and HDR. Chaker Larabi is a member of MPEG and JPEG committees. He serves as associate editor for the Springer SIVP journal and the SPIE/IS&T JEI. He is a member of CIE and IS&T, and a senior member of IEEE.

Libor Váša received his PhD from the University of West Bohemia, Czech Republic, in 2008. After a 3 year stay at Chemnitz University of Technology, Germany, he returned to University of West Bohemia in 2015, where he now works as Associate Professor. His scientific interests include compression of computer graphics data, such as mocap sequences or static and dynamic triangle mesh compression, together with the perceptual evaluation of the effect of lossy encoding of such data.